LB1
2565c

Columbia University

Contributions to Education

Teachers College Series

The Dartmouth College
Library
Has purchased this Book
from the Income of the
SANBORN LIBRARY FUND
Established by
EDWIN WEBSTER SANBORN
of the Class of 1878
In Memory of
EDWIN DAVID SANBORN
of the Class of 1832
for many years
Professor of Belles-Lettres
and sometime Librarian at
DARTMOUTH COLLEGE

Discarded by
Dartmouth College Library

Digitized by the Internet Archive
in 2026

https://archive.org/details/constructionvali0000marg

THE CONSTRUCTION AND VALIDATION OF AN ART TEST

BY

MARGARET McADORY, PH.D.

Mrs. Margaret (McAdory) Siceloff

TEACHERS COLLEGE, COLUMBIA UNIVERSITY
CONTRIBUTIONS TO EDUCATION, NO. 383

BUREAU OF PUBLICATIONS
Teachers College, Columbia University
NEW YORK CITY
1929

Discarded by
Dartmouth College Library

DARTMOUTH
COLLEGE
LIBRARY

COPYRIGHT, 1929, BY TEACHERS COLLEGE, COLUMBIA UNIVERSITY

700.7
S565c

FEB 25 1930
259020
Educ.

MARYMOUNT
COLLEGE
LIBRARY

ACKNOWLEDGMENTS

To Professor Frederick G. Bonser, of Teachers College, Columbia University, whose insight and continued interest have enabled the author to complete the research in its present form;

To Professor George J. Cox, Mr. Charles Martin, and Professor Sallie B. Tannahill, of Teachers College, for their criticisms which aided in establishing the graded series of test items;

To Professor Helen M. Walker and Professor Rudolf Pintner, of Teachers College, for valuable assistance in determining statistical techniques used in the construction of the test and for assistance in forming a usable instrument for testing in schools;

To Professor M. A. Bigelow, Director of the School of Practical Arts, Teachers College, Columbia University, President Samuel L. Duke, of State Teachers College, Harrisonburg, Virginia, Director Eugene Nifenecker, of the Bureau of Reference, Research and Statistics of the Board of Education of The City of New York, and Dr. N. R. Baker, former superintendent of schools of Jefferson County, Alabama, for the privilege of experimental testing;

And to the hundreds of judges who gave freely of their time in the preliminary checkings of the test items in order to establish the validity and reliability of the test, acknowledgment is gratefully expressed.

M. McA.

INTRODUCTION

The field of art education has made little use of the methods of testing results of teaching, the emphasis being on creative production with some parallel training in appreciation. Public school education in art is chiefly concerned with training in art appreciation. The McAdory Art Test * was devised to measure art appreciation, either as a group test or as an individual test, and to serve as an effective teaching aid by bringing to a focus the reflective judgment as applied to art values.

Each problem of the test comprises four illustrations which differ from each other in one art element (in a few cases more than one element was necessarily involved). The test is, therefore, a multiple response test which requires an ordered choice of the separate illustrations of each problem according to their relative merit. The scoring of such a test presents unique difficulties. No satisfactory precedent was found for the mechanics of scoring a multiple choice test that would take account of not only the first choice, but also the second, the third, and the fourth choices. The author used this art test as a basis for statistical research in determining reliable methods of scoring, and it is the purpose of this dissertation to present the statistical techniques and treatment applicable to a multiple choice test which requires ranking of responses in order of relative merit. The results are applicable in the construction and scoring of tests in any subjects involving ranked choices.

* The McAdory Art Test, published by the Bureau of Publications, Teachers College, Columbia University.

CONTENTS

CHAP.		PAGE
I.	Developing the Test	1
II.	Applying and Validating the Test	8
III.	Scoring the Test	20
IV.	Reliability and Validity	26
V.	How the Test May Be Used	30
VI.	Conclusions	34
	Bibliography	35

CHAPTER I

DEVELOPING THE TEST

BEGINNINGS

The construction and validation of the test for recognition of artistic merit had its beginning in 1923 in the rural schools of Alabama. There the preliminary tests were formulated and given to more than 2,000 elementary and junior high school children. Later some work was done on these preliminary tests in the State Teachers College at Harrisonburg, Virginia. The tests were started with the hope of proving that there is growth in the power to distinguish fine from poor examples and with the expectation that some of the characteristics of this growth could be learned and the results of art training measured.

Experimentation with the earlier tests yielded a method or technique by which they might be given to large groups in short periods of time. The most important result of the preliminary investigations was that the tests proved the need for a more complete analysis of art elements and for the inclusion of a broader range of subjects and common principles. The tests have been reorganized from time to time, and five years have been spent in selecting material and in making the tests valid and reliable.

ASSUMPTIONS

It was the purpose of these studies to construct a reliable instrument for the determining of the consensus of agreement or the order of preference of given subjects and art elements by experts, and for estimating or measuring the differences of agreement of groups and individuals. There are two underlying assumptions on which the items of the art test were constructed: first, that objective things can be ranked in order of artistic merit by a consensus of any group of people; second, that as far as any social group is concerned, its individual members can be ranked according to the degree of their agreement with the consensus adopted.

SUBJECTS CONSTITUTING THE TEST

The themes or subjects for the test were selected from common objects in everyday life as far as possible. Comparatively few people to-day can affect the production of the things they use or see, and even fewer are producers of art. All are consumers of art products and as such need to have their judgment developed in the selection of things they use or enjoy in the home or in activities outside.

The author aimed to make the subjects as impersonal as possible. Certain psychological considerations, however, affect the formation of individual judgments. For example, some persons might have a preference for a given arrangement of panels on a door because a friend may have had such a door. Similarly, prejudices for or against certain colors may have arisen from or been conditioned by subjective factors.

The test subjects comprise furniture, utensils, textiles, clothing, architecture, painting, and objects of other plastic and graphic arts. The testing material available is practically endless. Effort was made to select from a number of subjects those most commonly considered as having some relation to æsthetics. Such subjects are for convenience hereafter designated as test items.

SOURCE OF TEST ITEMS

Materials for the test items were selected from the following sources: current art magazines, trade magazines, department stores, art museums, books on special subjects, and articles used in homes.

Magazines supplied designs for automobiles, furniture, clothes, and rugs. Department stores were a source for items relating to dress accessories and clothes of men, women, and children. From art museums came designs in ceramics, silver, furniture, metal work, textiles, and paintings. Books on art subjects furnished designs in rugs, textiles, lace, printing and book-cover designs, sculpture, and furniture.

Material for hundreds of illustrations has been gathered from many sources. Mention should be made of photographs from the following museums: Metropolitan Museum of Art in New York, the Museum of Fine Arts in Boston, the Louvre in Paris, and the Cluny Museum in Paris. Any further acknowledgment of source material seems unnecessary, inasmuch as the test pre-

TABLE 1

LIST OF EXPERIMENTAL TESTS

Place	School System	Number of Persons Tested	Classification of Persons	Number of Test Items
Alabama, 1923-24	Jefferson County	2,000	Elementary and Junior High School Children in Rural Schools—Primary Grades	24 10
Virginia, 1924-25	State Teachers College at Harrisonburg, Va.	200 150	Students in State Teachers College at Harrisonburg, Va. — Children in Training School in City of Harrisonburg	27
Alabama, 1926-27	Jefferson County	300	Students in I and II Junior High School (Test was given 3 times: 900† checkings)	8
New York, 1927 May	Various Schools and Museums	20	Art Experts.....	60
Virginia, 1927 June-July	State Teachers College at Harrisonburg, Va.	205	Adults: Students and Faculty Members	60
Virginia, 1927 June-July	Training School	131	Children in Lower Grades Children in Higher Grades	50 60
New York, 1928 Spring Months*	Columbia University	3 10	Staff Members of Art Department, Teachers College, Columbia University; Graduate Art Students Engaged in Teaching	144 144
New York, 1928 May	Columbia University	40	Faculty Members and Artists	144
Virginia, 1928 June-July	State Teachers College at Harrisonburg, Va.	100	College Freshmen and Sophomores	72
New York, 1928 October	Columbia University	10 3	Class in Research in Elementary Education. Graduate Students. Staff of Metropolitan Art Museum Instructors	86 86
New York, 1928 November	Bronxville, N. Y.	20	Sixth Grade Bronxville Public School Children	72
New York, 1928 November	Teachers College, Columbia University	100 100	Art Experts; Art Students; Faculty Members of other than Fine Arts Department (25) Rechecking†. Art Students	86 86

TABLE 1 (*Continued*)

Place	School System	Number of Persons Tested	Classification of Persons	Number of Test Items
New York, 1928 January - February	Public School 165, New York City	115	Sixth Grade Children	60 Selected from 86 given in November

Total number of persons tested before May, 1928..... 3,006
Total number of persons tested since May, 1928..... 501

Grand total..... 3,507

*Tests given following the reorganization and redrawing of all tests and the addition of 84 test items.

†A total of 4,132 judgments which have been used in establishing the test units.

sentation of four different versions of each subject necessitated such changes as would differentiate them by steps of equal degrees of merit, and the mode of reproducton has often changed the material into forms quite different from the original.

Essential changes which might affect either the shape or color or values of dark and light of the original fine examples chosen for study as a basis for a test item were seldom made. The simplification for use as test material in some cases may, however, have made the reproduction less pleasing than the original.

NUMBER OF TEST ITEMS

Table 1 lists the experimental tests by date and school, giving the number and classification of persons taking the tests and the number of items in each test. The number of items varied from time to time. Originally there were 60 items, but in the spring of 1928 the test was redrawn and 84 items were added, when the study of statistical methods of treating the multiple choice test items was begun. Thus in May, 1928, the test was composed of 144 items. Seventy-two of those having the highest consensus order were used for testing students of the State Teachers College at Harrisonburg, Virginia. Of the remaining 72 items, 58 were discarded because of unsatisfactory consensus, and 14 were redrawn with the hope of increasing the indices of consensus. The test given in November, 1928, consisted of the 86 items. Some features of the study of the consensus of 100 judges in ranking

the four illustrations of each of these 86 items are presented in Tables 3 and 4. Sixty of the 86 items were selected for experimentation in comparing the reliability of various methods of scoring the test. The final selection of 60 test items was based on the criteria which are described in Chapter II. Sixty test items therefore form the art test and hereafter the term "art test" will refer to these 60 items.

ART ELEMENTS

The variables in the test items fall into three classifications of art elements. These elements were first described by the late Professor Arthur W. Dow, of Teachers College, Columbia University, in his early book, *The Theory and Practice of Teaching Art,* and more completely in *Composition,* a later work.

Line: an art element which gives definition to form, composing the contours and shapes of objects, and the breaking up of spaces into varied arrangements of parts of a whole.

Dark and Light: an art element which has to do with the use of amounts and distribution of black and white and all transitional values of light and dark. *Notan* is a single word used in *Composition* to denote this art element.

Color: an art element, derived from light, having to do with hues of red, yellow, green, blue, and purple, and all intermediate hues in various values of dark and light and of degrees of chroma.

The McAdory Art Test is organized to estimate the preferences for shapes and line arrangements, values of dark and light, and color—use of hue, value, and chroma.

LIMITING COMPARISONS TO A SINGLE VARIABLE

Wherever possible a single art element is involved in the comparison. For example, if it is a question of choice of shape of a bowl the decoration is eliminated or at least is the same on all objects presented for comparison. If it is a matter of color preference, then the shapes and values of dark and light remain constant. In some test items, however, it is necessary to vary two elements. For example, in the study of landscape architecture the houses and land structure remain the same but the arrangement and shapes of the shrubbery change in each illustration, and the amount of dark and light change with the altering in the arrangement of the shrubbery.

STRUCTURE OF TEST ITEMS

The test as a whole consists of 60 test items. Each test item is composed of four illustrations of the same subject or theme. The illustrations show four distinguishable degrees of merit represented by approximately equal steps.

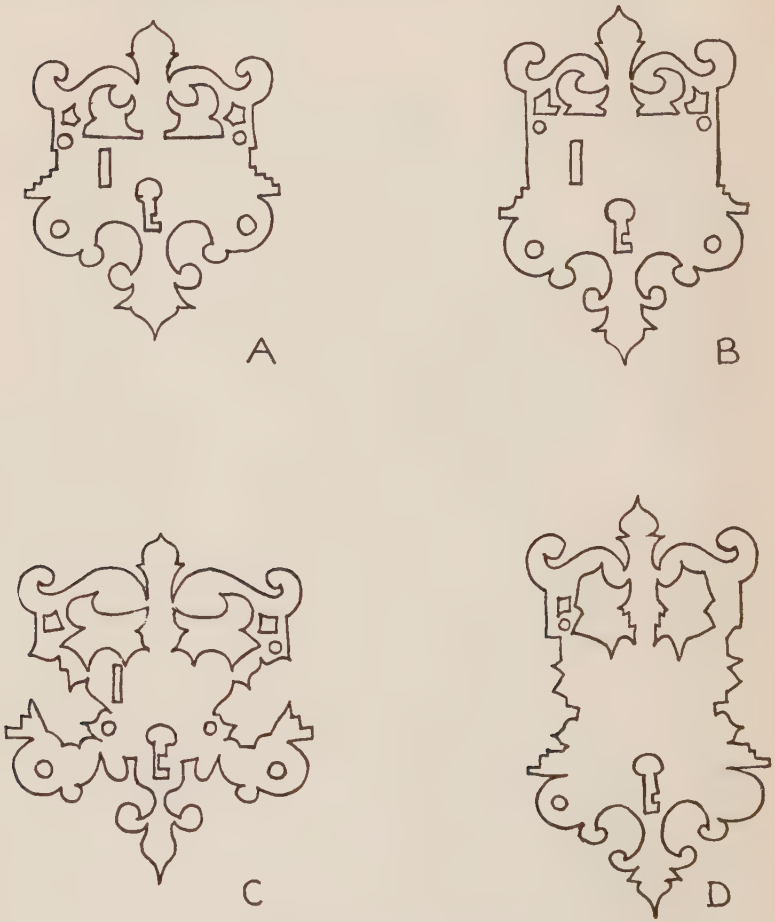

FIG. I. SAMPLE TEST ITEM

Figure I represents a test item. Illustration B is a drawing from an original escutcheon in the Museum of Fine Arts in Boston. It is beautiful in shape and proportion and is interesting in contour. Illustration A retains the simplicity of contour; the

change in proportion makes it less pleasing than B. The proportions of D are not bad and its use in certain instances would be satisfactory as far as the shape is concerned. However, the restless contour and the poorly related parts make it a little less pleasing than A. Illustration C comes fourth in the consensus of placement. The method of determining this order will be described and illustrated in subsequent pages.

Each test item presents for study one subject and as far as possible the consideration of a single element. For example line and shape were presented by Figure I. Each art element comes up with many different subjects and each subject presents every art element. The structure of the test provides for the alternation or rotation of groups of subjects and art elements. Table 3 gives a list of the test items. Items, 1, 2, 3, and 4 are of architecture and painting. Items 5, 6, 7, and 8 are of furniture and utensils. Items 9, 10, 11, and 12 are of textiles and clothing. Items 1 and 2, 5 and 6, 9 and 10 deal with line and shape. Items 3, 7, and 11 have to do with values of dark and light. Items 4, 8, and 12 deal with color.

RECURRENCE OF ELEMENTS AND SUBJECTS

Instead of being grouped in one section the test items are so arranged that similar subjects occur in different sections of the test. For example, the study of shapes of parts of chairs in relationship to the design of the complete chair will be found repeated from time to time. Similarly, the study of the right relationship of values of dark and light in rugs will be repeated. Because of this fact various parts of the test may be compared and correlated in order to determine the reliability of the test.

AVOIDANCE OF ARBITRARY ART RULES

No system of so-called rules of art has been followed in the construction of the test items. On the contrary, freedom has been reserved to draw upon the infinite possibilities of variation in the treatment of form and color where principles of contrast, of repetition, of subordination, of transition, and of symmetry are operative. Some of these uses may be fine and distinctive while others are commonplace or banal.

CHAPTER II

APPLYING AND VALIDATING THE TEST

For purposes of validation, the test of 86 items was given to 100 selected judges at Columbia University on November 21-22, 1928. In securing the judges, the author sent out more than 200 invitations to a non-selected group of persons listed in the *American Art Annual*. Those responding included the following: craftsmen, art buyers, painters, illustrators, art advisers of department stores, architects, members of museum staffs, art publishers and art editors, color specialists, art teachers, art supervisors, directors of art, college professors, art collectors, and advanced art students. Their reactions to the test furnished the material for the conclusions here given.

SELECTION OF JUDGES TO DETERMINE DEGREES OF MERIT

In the selection of judges the following principles were observed:

The experts should be chosen from a broad range of fields having to do with art and art appreciation.

The number should be large enough and the agreement satisfactory.

The range should include the following groups:

1. Art producers: (*a*) architects, (*b*) painters, (*c*) craft workers and manufacturers.

2. Art teachers: (*a*) colleges, (*b*) art schools, (*c*) high schools, (*d*) elementary schools.

3. Critics: (*a*) museum curators and lecturers, (*b*) art editors, (*c*) writers, (*d*) collectors.

4. Art buyers and trained workers from department stores or specialty shops.

5. Competent lay critics: people of college education who have traveled and have a general interest in art, but who are not engaged in producing, teaching, or selling.

METHOD OF DETERMINING CONSENSUS OF JUDGES

Table 2 presents a treatment of the reactions of 30 judges in order to determine the consensus order of placement. Twenty-three of the judges gave B first place; one judge gave B third place and A fourth place.

Thus, twenty-four, or 80 per cent, of the judges placed B as better than A.

Twenty-one, or 70 per cent, of the judges placed A before D.

Twenty-one, or 70 per cent, of the judges preferred D to C.

The order of placement resulting is: B A D C. This order shall be referred to as the consensus, and the percentage numbers, 80, 70, 70, as the indices of the consensus.

TABLE 2

EVALUATION OF FOUR ILLUSTRATIONS CONSTITUTING A TEST ITEM AS GIVEN IN FIGURE I

Order of Choice				Number of Judges
A	B	C	D	
2	1*	4	3	9
3	1*	4	2	8 } 23*
2	1*	3	4	6
1	2	4	3	3
4	3	2	1	1
1	3	2	4	3
Total.....				30

* Twenty-three judges placed B first.

SATISFACTORY MINIMUM INDEX OF CONSENSUS

The number of judges upon whose decision the final acceptance of validity was based was 100, with consensus indices of not less than 64 for every item retained in the test. This was considered a satisfactory minimum index of consensus for 100 judges, because then the probability that a second jury of 100 similarly qualified judges would change a given consensus order with minimum indices, 64, 64, 64, is less than one chance in a hundred. This probability was obtained in the following way. Let us assume

that A has been rated above B with 64 per cent of agreement. The standard deviation of this per cent is $\sigma = \sqrt{\frac{(.64)(.36)}{100}} = .048$. The probability that another jury similar to this one would rate B above A is the probability that this per cent of agreement should fall below .50. But .64—.50 = .14, which is 2.92 times the standard deviation; and in Pearson's *Tables for Statisticians and Biometricians,* Table II, page 5, we find that the area under the normal curve up to the ordinate corresponding to $x = 2.92\sigma$ is .99825 of the entire area under the curve; thus leaving only .00175 of the area beyond this ordinate. Hence .00175 is the probability that a new jury would rate B above A. This is less than two chances in a thousand.

This probability applies only to the interchange of two specified adjacent letters in a consensus. Obviously, an equal or smaller probability applies to the interchange of two non-adjacent letters. Now the 24 possible orders of the four letters, A, B, C, D, can be obtained from any one such order by interchanges of pairs of letters. Suppose, for example, that a given consensus is B A D C. Interchanging A and B gives A B D C. There are 6 orders, each of which is obtained by one interchange of two letters; there are 9 orders, each obtained by two consecutive interchanges (for example, interchange A and C in B A D C, obtaining B C D A, then interchange C and D, obtaining B D C A); and the remaining 8 of the 24 orders result from 3 successive interchanges each.

But the probability that a new jury of 100 judges would interchange 2 pairs of letters in a given consensus of a previous jury is equal to or less than the product .00175 × .00175, or about .000003; and the total for 9 such orders may be ignored, as may the 8 orders involving 3 interchanges each. There remain only the 6 orders involving a single interchange each, and the probability that one or another of these will be voted by the new jury is equal to or less than 6 times .00175, or .0105. We conclude, therefore, that with 100 judges, a minimum value of .64 for consensus indices sufficiently protects the validity of the test.

FINDINGS OF THE JUDGES

When a comparison of percentages of agreement and of deviations from group agreements was made it was found that art

specialists make higher scores than art students or members of the miscellaneous group of cultured persons. The art specialists have fewer deviations on 42 of the test items; the advanced art students have fewer deviations on 23 items; the miscellaneous group have fewer deviations on 21 test items.

Tables 3 and 4 give the complete record of the judgments of the 100 selected judges on 86 test items. The two tables are printed on facing pages so that they can be read across and all the data of each item thereby read at once. Table 3 includes: (*a*) subjects of each item, (*b*) art elements, (*c*) consensus order, (*d*) indices of consensus, (*e*) percentage of deviation from consensus. Table 4 is a similar analysis by groups among the 100 judges: (*x*) specialists; (*y*) art students; and (*z*) miscellaneous group.

Table 3 is to be read thus: Item No. 1 has as its subject, lettering. The art element involved is shape. The consensus is given in column (*c*): A first, C second, D third, and B fourth. The indices of this consensus show that 95 per cent placed A above C, that 90 per cent placed C above D and 94 per cent placed D above B. The per cent of deviation from this order is 8.0. On the opposite page, Table 4 is to be read as a continuation of Table 3: The order of choice for Item 1 is the same for the 34 art specialists. The indices of consensus are 91, 97, 94. The per cent of deviation is 7.3. For the 30 art students the order of choice is the same. The indices of consensus are 97, 87, 93. The per cent of deviation is 9.2. The 36 persons of the miscellaneous group have made the same order of choice. Their indices of consensus are 97, 94, 94. The per cent of deviation is 9.0.

ANALYSIS OF RESULTS: PERCENTAGE OF AGREEMENT

An analysis of the indices of consensus shows that 68 out of the 86 test items received 60 per cent or greater agreement in the placement of the four illustrations. In all except 11 test items the three groups comprising the 100 judges agreed on the order of placement as shown in the tables. The exceptions are test items 3, 11, 13, 16, 24, 32, 44, 49, 54, 59, and 74. The test items receiving less than 60 per cent agreement were the last ten of those just given, and in addition items 14, 15, 33, 75, 76, 79, 80, 83.

A further study shows that 61 test items received an agreement of at least 64 per cent for the placement of four illustrations. from 60 to 64 per cent are numbers 6, 10, 21, 45, 51, 66, and 77.

TABLE 3

CONSENSUS JUDGMENT OF 100 JUDGES ON 86 TEST TIMES

Item	(*a*) Subject of Test Items	(*b*) Art Element	(*c*) Consensus Order	(*d*) Indices of Consensus	(*e*) Per Cent of Deviation from Consensus
1	Lettering	Shape	ACDB	95,90,94	8.0
2	Ships	Line Arrangement	CADB	73,84,93	17.2
3	Painted Houses	Dark and Light	DBAC	65,72,64	40.0
4	Paintings	Color	CBDA	85,90,81	14.5
5	Bowls	Shape	DBAC	69,80,87	21.0
6	Teapots	Shape	BCAD	63,86,81	17.5
7	Automobiles	Dark and Light	BCDA	89,91,76	12.0
8	Chests	Color	DABC	82,66,73	25.0
9	Hats	Shape	ACDB	88,79,81	15.3
10	Dresses	Line Arrangement	CDBA	61,86,73	22.5
11	Dresses	Dark and Light	DABC	66,90,52	24.5
12	Sweaters	Color	ACDB	89,74,74	18.5
13	Printing	Line Arrangement	DBAC	80,87,58	26.0
14	Houses	Shape	DBAC	85,58,75	27.7
15	Locks	Line Arrangement	BCAD	81,81,56	25.0
16	Painting	Color	CBDA	51,72,81	29.3
17	Chairs	Shape	CADB	82,75,88	14.2
18	Plates	Line Arrangement	ABCD	66,72,78	27.2
19	Cabinets	Dark and Light	DBAC	81,85,93	12.5
20	Plates	Dark and Light	BDAC	89,79,83	15.0
21	Handkerchiefs	Line Arrangement	BACD	87,78,60	31.2
22	Boys' Suits	Dark and Light	DBAC	75,88,89	16.5
23	Child's Dresses	Color	ACBD	86,75,81	17.5
24	Textile	Dark and Light	ACBD	63,50,73	36.2
25	Houses	Arrangement	DBAC	84,88,86	11.2
26	Gates	Line Arrangement	DCAB	73,87,71	16.0
27	Trees with house	Dark and Light	ACDB	92,80,84	16.7
28	Houses	Color	CABD	89,86,92	12.2
29	Pitchers	Shape	BACD	88,93,95	7.7
30	Tables	Shape	ABCD	87,95,98	8.0
31	Tables	Dark and Light	CDAB	94,89,90	8.7
32	Chests	Painted Color	BDAC	51,74,90	26.7
33	Coats	Shape	BACD	56,99,82	18.
34	Dresses	Dark and Light	ADBC	97,91,65	14.7
35	Lace	Dark and Light	CBAD	71,87,89	16.
36	Dresses	Color	ACDB	85,78,71	24.2
37	Sculpture	Shape	BDCA	84,87,78	22.7
38	Doors	Shape	CADB	69,86,71	21.7
39	Paintings	Color	BACD	69,67,79	29.3
40	Paintings	Color	DBCA	68,79,81	24.0
41	Chairs	Line Arrangement	CABD	86,89,83	11.0
42	Chalice	Shape	CABD	84,78,81	19.2
43	Automobiles	Dark and Light	ABDC	97,68,77	17.2
44	Chests	Dark and Light	DCAB	93,51,90	21.0
45	Rugs	Dark and Light	BADC	67,76,60	36.7
46	Dresses	Shape	ABCD	78,74,98	16.0

(*Continued on page 14*)

TABLE 4

REPORT OF SEPARATE GROUP CHECKINGS OF 86 TEST ITEMS

Item	(x) 34 Specialists			(y) 30 Art Students			(z) 36 Miscellaneous		
	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (c) of Table 3	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (c) of Table 3	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (c) of Table 3
1		91,97,94	7.3		97,87,93	9.2		97,94,94	9.0
2		74,82,94	16.9		70,80,90	18.3		72,89,94	16.7
3		65,71,71	33.	DBCA	67,67,71	57.		64,72,72	32.
4		85,91,91	9.6		77,80,90	20.0		94,100,83	14.6
5		68,76,94	19.6		63,90,77	23.3		75,75,89	20.1
6		65,82,88	15.5		77,77,70	21.7		50,97,83	16.0
7		85,94,65	14.7		93,93,83	11.7		89,86,80	9.7
8		85,74,74	21.3		80,63,77	26.7		80,61,70	27.1
9		85,88,85	12.5		80,80,83	15.8		97,70,75	15.9
10		65,76,79	22.7		60,90,80	20.8		58,92,61	23.6
11	DACB	71,100,63	25.7		67,90,57	22.5		61,92,56	25.6
12		91,70,75	16.2		93,77,63	17.5		86,72,78	21.5
13		85,78,64	25.7	DBCA	90,87,63	27.0		67,92,67	25.0
14		82,50,76	30.1		80,67,77	24.1		92,58,72	28.4
15		79,82,56	26.4		97,83,60	20.		70,78,53	27.7
16	BCDA	56,65,79	27.9		50,77,80	29.1		58,61,89	31.2
17		82,65,91	17.3		80,83,80	17.5		100,82,94	8.3
18		53,76,79	26.2		77,77,73	30.0		70,64,80	25.7
19		93,93,77	10.3		90,87,87	12.5		72,86,94	14.6
20		91,85,85	11.0		83,63,87	20.0		92,86,78	14.6
21		85,79,53	30.9		87,73,63	29.1		89,80,64	30.9
22		74,82,88	19.1		80,90,93	14.2		72,92,86	16.0
23		76,76,79	17.6		87,83,70	20.		94,67,92	15.3
24		65,59,56	40.5	ABCD	63,53,87	33.3	ABCD	61,47,67	34.6
25		74,85,85	10.6		87,87,87	15.0		92,92,86	9.0
26		71,97,74	13.9		70,80,77	16.7		78,83,64	17.3
27		94,85,85	19.0		93,80,80	15.8		94,75,80	15.3
28		85,82,94	12.5		93,90,93	10.0		89,86,89	13.8
29		88,100,88	5.8		93,97,97	4.1		83,83,100	12.5
30		94,97,97	5.1		85,90,100	12.5		83,97,97	7.0
31		94,88,91	8.8		90,90,90	10.0		92,89,89	7.6
32		53,74,97	24.2	DBAC	57,70,87	29.1		56,67,86	27.1
33		59,100,94	14.3		50,97,97	17.5		58,100,58	22.9
34		97,94,56	18.4		97,93,63	15.0		97,86,75	11.1
35		76,97,94	9.5		57,83,90	19.1		78,80,83	19.4
36		85,76,59	30.8		90,87,77	26.6		80,72,78	29.8
37		76,97,88	14.7		83,80,83	12.5		89,83,64	28.5
38		74,88,76	19.9		57,90,63	25.		75,80,72	20.8
39		76,59,76	30.8		67,73,87	26.6		64,70,75	29.8
40		62,82,85	25.0		67,77,83	20.8		75,78,75	22.5
41		91,94,79	9.6		87,90,83	10.0		80,83,86	13.2
42		79,79,79	21.3		87,67,80	20.0		86,86,83	16.7
43		97,68,76	15.4		93,73,77	15.8		97,64,78	20.1
44		91,56,88	19.8		93,50,87	23.3	DACB	94,53,94	20.1
45		76,71,56	43.4		57,87,60	38.3		67,72,64	31.9
46		85,74,97	14.0		87,70,100	14.2		64,78,97	19.4

(*Continued on page 15*)

TABLE 3

(*Continued from page 12*)

Item	(*a*) Subject of Test Items	(*b*) Art Element	(*c*) Consensus Order	(*d*) Indices of Consensus	(*e*) Per Cent of Deviation from Consensus
47	Dresses	Dark and Light	BACD	75,95,95	9.2
48	Sweaters	Color	BACD	80,75,86	20.0
49	Buildings	Shape	BDAC	55,91,91	23.0
50	Buildings	Line Arrangement	CABD	89,90,92	8.7
51	Books	Dark and Light	DCAB	61,81,86	23.0
52	Portraits	Dark and Light	ADCB	72,80,69	23.
53	Pitchers	Line Arrangement	BDAC	65,77,84	27.0
54	Spoons	Shape	ACDB	98,92,53	18.5
55	Mugs	Dark and Light	ACDB	82,68,88	19.5
56	Plates	Color	BADC	75,85,72	26.2
57	Tile	Dark and Light	ADBC	79,89,82	22.0
58	Room Arrangements	Dark and Light	DABC	79,77,85	17.2
59	Rugs	Dark and Light	DBAC	84,53,90	24.0
60	Rugs	Color	ADBC	68,87,86	17.5
61	Lettering	Shape	CDAB	87,92,72	13.2
62	Landscape	Arrangement & Shape	DBAC	71,75,73	32.0
63	Painting	Dark and Light	CDAB	64,72,73	29.5
64	Painting	Dark and Light	ABCD	72,86,77	19.6
65	Highboys	Shape	BCDA	88,73,79	22.2
66	Davenports	Shape	BADC	89,98,60	13.2
67	Chairs	Shape	BDAC	87,79,98	11.0
68	Davenports	Color	ABCD	65,85,82	20.7
69	Blouses	Line Arrangement	CADB	69,80,85	23.3
70	Borders	Line Arrangement	CABD	80,88,90	14.0
71	Rugs	Dark and Light	BADC	90,84,76	19.7
72	Child's Dresses	Color	ABCD	84,75,79	19.2
73	Room Interiors	Dark and Light	CBDA	69,77,71	27.2
74	Buildings	Shape	BDAC	72,62,54	47.0
75	Room Arrangements	Dark and Light	CDAB	57,85,78	22.7
76	Candlesticks	Shape	BACD	73,98,55	20.7
77	Rugs	Color	DBAC	63,68,86	34.7
78	Church Towers	Shape	CABD	68,84,93	23.0
79	Metal Work	Line Arrangement	CBAD	67,58,91	32.7
80	Bowls	Color	ABDC	64,56,65	38.2
81	Textile Designs	Dark and Light	CBAD	70,65,92	23.7
82	Page of Printing	Line Arrangement	ABCD	86,93,78	11.7
83	Davenport	Color	BCAD	76,56,68	35.5
84	Vases of Flowers	Color	ADCB	74,74,64	33.2
85	Gravestones	Shape	BACD	68,83,69	27.5
86	Ends Papers	Dark and Light	DCAB	72,76,90	19.5

NOTE: The final selection of 60 test items from these 86, made by applying criteria of validity and appreciative preference, is found in Table 6, page 18.

TABLE 4

(*Continued from page 13*)

Item	(x) 34 Specialists			(y) 30 Art Students			(z) 36 Miscellaneous		
	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (*c*) of Table 3	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (*c*) of Table 3	(1) Consensus	(2) Indices	(3) Per Cent of Deviation from (*c*) of Table 3
47		71,97,100	9.1		77,93,97	8.1		78,94,89	10.4
48		85,70,97	19.		83,83,90	15.0		72,72,72	26.4
49		56,94,97	19.8	DBAC	63,63,80	38.3		70,94,94	13.2
50		88,100,91	9.6		90,93,97	10.0		83,94,86	8.3
51		68,85,85	16.9		60,83,80	26.7		56,75,92	25.7
52		71,88,97	18.7		80,77,73	24.1		67,75,67	29.8
53		79,76,85	21.3		67,80,90	21.7		50,75,78	36.8
54		100,94,65	14.0		100,93,57	15.8	ACBD	94,58,61	21.5
55		94,78,83	14.		87,57,90	21.7		67,78,83	23.
56		65,79,82	36.4		80,83,80	15.0		80,92,56	25.7
57		74,91,91	23.5		87,90,77	20.8		78,86,78	21.5
58		79,76,80	14.7		87,77,77	16.7		72,78,83	20.1
59		85,56,88	25.7		67,60,90	25.0	DABC	97,56,56	21.5
60		68,82,82	18.4		70,87,80	20.8		67,92,94	13.9
61		88,94,68	14.3		80,93,80	13.3		92,89,70	11.8
62		74,79,71	27.9		70,80,83	32.5		70,67,67	35.4
63		59,79,71	32.3		60,77,70	17.5		72,53,78	36.8
64		76,91,82	15.4		60,87,63	25.0		78,80,83	18.1
65		85,71,85	22.1		77,67,87	21.7		72,80,67	23.6
66		91,100,59	13.2		90,97,60	12.5		86,97,61	13.9
67		97,85,97	7.3		93,90,97	6.7		72,64,100	18.1
68		68,88,82	19.7		60,83,80	22.5		67,83,83	20.1
69		65,82,91	18.4		73,90,87	23.3		70,70,78	27.8
70		65,79,91	20.6		93,90,90	9.1		86,94,89	11.8
71		94,79,79	16.8		93,97,63	20.0		83,78,83	22.2
72		88,76,79	15.4		93,63,67	23.3		72,83,83	19.6
73		65,74,71	29.4		70,67,77	30.8		72,89,67	22.2
74		82,74,76	37.1		63,50,53	45.8	BDCA	75,67,67	56.9
75		50,91,85	20.6		57,83,77	25.0		64,80,72	22.9
76		76,100,53	18.7		70,100,63	20.8		72,94,50	22.2
77		68,59,91	40.4		57,73,80	35.8		64,72,86	28.1
78		85,79,94	21.3		63,90,97	29.1		56,83,89	17.5
79		74,56,100	33.1		53,67,87	30.8		72,53,86	34.
80		76,53,59	36.0		63,57,63	37.5		53,58,72	40.6
81		88,74,100	16.7		73,70,97	19.1		50,53,80	36.1
82		97,91,88	5.9		100,93,80	8.3		64,92,67	20.1
83		74,56,71	35.3		80,57,67	36.7		75,56,67	34.7
84		82,82,53	28.7		77,77,80	35.7		64,64,61	35.4
85		76,88,76	17.6		67,73,73	28.1		61,86,58	35.4
86		79,71,97	14.0		80,73,87	20.7		58,83,86	23.5

NOTE: When the order of consensus is the same for the smaller groups as the order of consensus or placement for the whole groups [(*c*) of Table 3], the space has been left blank. Differences in order of agreement have been indicated.

The numbers of test items receiving agreement of 65 per cent or more are: 1, 2, 4, 5, 7, 8, 9, 12, 17, 18, 19, 20, 22, 23, 25, 26, 27, 28, 29, 30, 31, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 46, 47, 48, 50, 52, 53, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 78, 81, 82, 84, 85, 86. In these items the agreement is considered satisfactory, because there is only about one chance in 100 that another similar set of judges would reverse the order of merit of the illustrations.

METHOD OF DETERMINING DEVIATION PERCENTAGES

Table 5 shows the method of determining the percentages of deviation, illustrating with data for Item 18. The plan of the table will be clear if we follow across the second line showing the order 1243, with its 2 points of deviation from the consensus order, which was voted by 1 art specialist, 2 art students, and 4 persons of the miscellaneous group. There the 3 is one point out of order, as is also the 4; hence there are 2 points deviation. Likewise in the fifth line, with order 2314, the 4 is correct but the 2 and 3 are each one place removed from the proper placement, and the 1 two places removed; hence this ranking would have 4 points deviation. It was possible for each person to deviate 8 points from the consensus on one item. Thus for the 34 art specialists the total number of possible deviations is 272. It will be seen from the table that in the group of art specialists,

11 made 0 deviations.
† 12 made 2 deviations each, or a total of 24
‡ 9 made 4 deviations each, or a total of 36
§ 2 made 6 deviations each, or a total of 12
Total 72

The total number of deviations is 72, or 26.4 per cent of the number possible. In the same way the deviations of art students are found to be 30 per cent, and of the miscellaneous group 27.7 per cent. The total possible number of deviations for 100 people is 800. The 100 judges actually made 218 deviations or 27.2 per cent of the total possible number.

CRITERIA FOR THE SELECTION OF TEST ITEMS

Criteria observed for the selection of test items were as follows: first, that the percentage of agreement about order of placement of four illustrations be satisfactory; second, that the items be

TABLE 5

METHOD OF DETERMINING DEVIATIONS*

Various Orders Voted by 30 Judges	Number of Points of Deviation	Number of Judges Arranging Choices in a Given Order		
		34 Art Specialists	30 Art Students	36 Miscellaneous
ABCD				
1 2 3 4 (consensus)	0	11	9	10
1 2 4 3	2	1	2	4
1 3 2 4 †	2	2	3	7
2 1 3 4	2	9	5	5
2 3 1 4	4	3	3	3
2 1 4 3 ‡	4	5	5	3
3 2 1 4	4	1	..	2
3 1 2 4	4	..	1	1
4 2 1 3	6	1	..	..
2 4 1 3 §	6	..	1	1
3 1 4 2	6	..	1	..
2 4 3 1	6	1	..	..

*For Item 18, checkings of 100 judges.

chosen with reference to percentage of deviations from the correct order of placement; third, that an approximate balance of subjects constituting the test be retained and that art elements of line and shape, and dark and light, and color be well represented; and fourth, that, among the items satisfying the above criteria, the preference be given those test items liked best by 50 of the judges who upon request indicated preferences.

The first criterion for the final selection of test items is that the indices of consensus on each item should be 64 or more. The second criterion, percentage of deviations, is very closely related to the first. When the test items are ranked according to the per cent of deviations from an agreed order, some of the items not included on the basis of the first criterion, because the indices of consensus have been under 64 per cent, will nevertheless be included because they have a very low percentage of deviation on the item as a whole. Item 18, for instance, as illustrated in Table 5, has 27.2 per cent deviation. Item 15 has only 25 per cent deviation and has

a satisfactory consensus of agreement on all places. Item 29 had the lowest percentage of deviations, 7.7. Items 1, 30, 31, 50 come next with 8 per cent, and so on, up to 40 per cent deviations on test item 3. The consensus of agreement on this item is 65, 72, 64, so that this item satisfies the first criterion, but is rejected because of the high percentage of deviations. Item 62 should be included in a final selection on a basis of consensus, having 71, 75, 73, but on a basis of percentage deviations it was excluded on account of its very high deviation, 32 per cent. On the other hand, test item 66 would be excluded on a basis of consensus,

TABLE 6

FINAL SELECTION OF TEST ITEMS

PART I—30 ITEMS	PART II—30 ITEMS
1 Lettering	42 Chalice
2 Ships	43 Automobiles
4 Painting	45 Rugs
5 Bowls	46 Dresses
7 Automobiles	47 Dresses
8 Hats	48 Sweaters
10 Dresses	50 Buildings
12 Sweaters	51 Books
17 Chairs	52 Portraits
18 Plates	53 Pitchers
19 Cabinets	55 Mugs
20 Plates	56 Plates
21 Handkerchiefs	57 Tile
22 Boy's Suits	58 Room Arrangement
23 Child's Dresses	60 Rugs
25 Houses	63 Painting
26 Gates	66 Davenports
27 Landscape Arrangement	67 Chairs
28 Houses	65 High boys
30 Tables	64 Painting
31 Tables	69 Blouses
34 Dresses	70 Border Designs
37 Sculpture	71 Rugs
38 Doors	72 Child's Dresses
39 Painting	78 Church Towers
40 Painting	81 Textiles
41 Chairs	82 Pages of Printing
85 Gravestones	86 End Papers

having 89, 98, 60, but it was included because the deviation was only 13 per cent.

The following groups of the test items show the order of preferences of 50 of the 100 judges who checked the items. The items of Group I were liked best, items of Group II next, and so on to Group VI which was liked least.

Group I: 9, 20, 67, 75, 79, 81, 82—liked best.

Group II: 2, 7, 13, 15, 16, 18, 22, 30, 42, 46, 54, 57, 62, 64, 65, 69, 74, 76, 84—liked fairly well.

Group III: 4, 17, 21, 27, 28, 33, 34, 35, 38, 48, 51, 52, 53, 60, 70, 71, 77, 80, 85.

Group IV: 11, 14, 23, 24, 25, 26, 29, 31, 33, 39, 40, 43, 45, 49, 50, 55, 56, 59, 63, 68, 72, 73, 78, 83, 86.

Group V: 3, 6, 8, 19, 36, 44, 47, 58—little liked.

Group VI: 1, 12, 32, 61, 66—not liked.

Those items not liked or little liked were discarded, with the exception of seven retained for special reasons.

FINAL SELECTION OF TEST ITEMS

With the information obtained by the grouping described above, sixty test items were selected. The test is divided in two parts of 30 items each as shown in Table 6.

CHAPTER III

SCORING THE TEST

For scoring multiple choice tests requiring ranked order of choices, no satisfactory technique had been developed. Experimentation was made with every form of scoring which the author

TABLE 7

RELIABILITY OF METHODS OF SCORING

(90 children; 10 test items)

Method of Scoring Test	Self r
1. Rating 1st choices: 1st is put 1st, 10 points 1st is put 2nd, 7 points 1st is put 3rd, 3 points	.137
2. Rating 1st choice correct, 1 point	.26
3. Rating 1, 2, 3, 4 all correct, 10 points 1 and 4 as correct, 5 points Either 1 or 4 as correct, 5 points	.21
4. No deviations from the correct order, 20 2 deviations from the correct order, 15 4 deviations from the correct order, 10 6 deviations from the correct order, 5	.23
5. Rating 1 and 4 correct, 10 points 1 or 4 correct, 5 points	.32
6. No deviations from correct order, 10 points 2 deviations from correct order, 9 points 4 deviations from correct order, 7 points 6 deviations from correct order, 3 points 8 deviations from correct order, 1 point; but correct order reversed, 0 points	.43
7. Rating one point for each correct placement either for 1, 2, 3, or 4*	.45

* The last method is the most reliable and is a very easy method of scoring. The Spearman-Brown formula applied to results for 10 problems gives $r_{1\,II} = .60$.

TABLE 8

METHODS OF SCORING AND RELIABILITY COEFFICIENTS OF TESTS GIVEN TO 100 FRESHMAN AND SOPHOMORE STUDENTS OF HARRISONBURG

(72 items: 36 in first half, 36 in second half)

Method of Scoring for each Item	Self r
1. No deviations (4 correct placements*), 1 point	.48
2. First choice correct, 1 point	.716
3. 1st and 4th correct, 1 point	.65
4. 1st and 4th correct 2 points 1st or 4th correct 1 point	.78
5. 1st and 4th correct, 3 points 1st or 4th correct, 1 point	.71
6. Total number of deviations on problems of 1st half correlated with deviations on problems of 2nd half	.71
7. Weighted deviations, 0 deviation 10 points Weighted deviations, 2 deviations 9 points Weighted deviations, 4 deviations 7 points Weighted deviations, 6 deviations 3 points Weighted deviations, 8 deviations 0 points	.63
8. 0 deviation, 2 points 2 deviations, 1 point	.81
9. 0 deviation 2 points 2 deviations 1 point 4 deviations 1 point	.83
10. 0 deviation 1 point 2 deviations 1 point	.77
11. 0 deviation, 2 points 2 deviations, 1 point 8 deviations, − 1 point	.70
12. 0 deviation, 2 points 2 deviations, 1 point 6 deviations, − 1 point 8 deviations, − 1 point	.83
13. 0 deviation, 2 points 2 deviations, 1 point 6 deviations, − 1 point	.77
14. 0 deviation, 2 points 2 deviations, 1 point 6 deviations, − 1 point 8 deviations, − 2 points	.78
15. 0 deviation, 5 points 2 deviations, 1 point 4 deviations, 1 point	.70
16. 0 deviation, 3 points 2 deviations, 1 point	.73
17. 0 deviation, 3 points 2 deviations, 1 point 8 deviations, − 1 point	.76

* It is assumed that a placement is correct when it is in agreement with the consensus order.

TABLE 8 (*Continued*)

Method of Scoring	Self *r*
18. 0 deviation, 3 points 2 deviation, 1 point 4 deviations, 1 point	.70
19. 0 deviation, 3 points 2 deviations, 1 point 4 deviations, 1 point 8 deviations, − 1 point	.76
20. 0 deviation, 4 points 2 deviations, 1 point	.63
†21. 1 point for each correct placement: 1st, 1 point; 2nd, 1 point; 3rd, 1 point; 4th, 1 point	.869

† $r_{1\,II}$ (72 Items) = .93.

TABLE 9

COMPARISON OF RELIABILITY COEFFICIENTS DERIVED BY DIFFERENT METHODS OF SCORING CHECKINGS OF 100 JUDGES

(60 test items)

Method of Grading Methods numbered as in Table 8	Self *r*	$r_{1\,II}$	S. D.			Mean	S. D. of Mean
			First Half	Second Half	Whole		
2. 1st choice correct, 1 point	.537	.69	3	3.9	6.08	44.1	.6
8. 0 deviations, 2 points 2 deviations, 1 point	.85	.92	8.5	9.00	17.5	69.35	1.75
21. 1st choice correct, 1 point 2nd choice correct, 1 point 3rd choice correct, 1 point 4th choice correct, 1 point	.86	.93	12.	14.	25.	160.9	2.5

could devise (twenty-six in number) and the reliability of the results of the different methods was tested.

The most extensive of these studies was developed in the State Teachers College at Harrisonburg, Virginia. There a preliminary study was made of the papers of 90 children, grades 3 to 7 inclusive. Table 7 shows seven different methods of scoring ten test items chosen from sixty and gives the coefficient of reliability for each method.

A second study was made of the checkings of a test of 72 items by 100 college students as reported in Table 8. The third study was made of the papers of 100 judges, using 60 test items. The results are given in Table 9. The findings of each of these studies are included in this paper. An analysis of the second study, which was the most exhaustive, will give the key to the fourth and is sufficient for the others as well. Table 8 shows twenty-one different self-correlations that were made in the search for a method which would give highest reliability from the given number of items.

The first five methods of scoring need no explanation, as they are easily understood:

1. Four placements of a test item correct, one point.
2. The first choice correct, one point.
3. The first and fourth choices both correct, one point.
4. The first and fourth choices correct, two points; first or fourth choice correct, one point.
5. First and fourth choices correct, three points; first or fourth choice correct, one point.

Beginning with the sixth method of scoring and continuing through the twentieth method, the number of deviations from the consensus must be known. Tables 5 and 10 show how the deviations are counted. A helpful device in counting the scores was a sheet of transparent paper on which all the correct placements were marked for each of the items. This was laid over the test paper in such a way that the numbers could be easily compared and the total number of correct placements readily counted.

Table 10 illustrates the method of counting deviations. In the first line, 3241 is the consensus order and it has no point of deviation. In the second line, 2341 has 2 points of deviation from the consensus; 2 is a deviate from 3 and 3 is a deviate from 2.

Thus for each person taking the test, each test item was ex-

amined, and the exact number of deviations found for each half of the test as given under 6 of Table 8. Values were assigned as shown in methods of scoring 7 through 20 of Table 8.

TABLE 10

METHOD OF COUNTING DEVIATIONS

Order	Number of Deviations
ABCD (consensus)	
3 2 4 1	0
2 3 4 1	2
3 1 4 2	2
4 2 3 1	2
3 4 2 1	4
3 2 1 4	6
1 3 2 4	8
2 3 1 4	8

Table 11 shows the 21st method of scoring: one point for each correct placement. The correct order is understood to be 3241 (or D B A C). Thus 4 would be the highest score per item.

TABLE 11

METHOD OF SCORING BY COUNTING CORRECT PLACEMENTS

Order	Score
ABCD	
3 2 4 1 (consensus)	4
3 1 4 2	2
4 2 3 1	2
2 3 4 1	2
2 4 3 1	1
2 3 1 4	0

On grounds of reliability and ease of scoring, methods 2, 8, and 21 were selected as the best. These three methods yield approximately the same reliability with 100 college freshmen, Table 8, as with 100 judges, Table 9. Table 9 shows certain other statistical data on the reliability of these three methods.

The 21st method of scoring is objective, is easy to apply, and is rapid. It gives a high coefficient of reliability when two parts of the test are correlated; namely, .869 for 36 test items, and .93 for 72 test items (See Table 8). A comparison of this method of scoring with other methods is given in Table 7, Table 8, and Table 9.

CHAPTER IV

RELIABILITY AND VALIDITY

The reliability and validity of the test were computed by comparing scores for different groups and for sections of the same group. Table 12 gives the coefficients of reliability for 100 art students, 100 sixth grade children, and 25 advanced art students. Figure II exhibits comparative group scores.

For the purpose of studying the validity of the test, the 100 judges were divided into three groups. Figure II shows the over-

TABLE 12

RELIABILITY OF A TEST OF SIXTY ITEMS FOR SELECTED GROUPS

(100 Fine Arts Students, 100 Sixth Grade Children, 25 Advanced College Art Students)

Group*	Self r	$r_{I\,II}$	Standard Deviation			Mean	S. D. of Mean
			First Half	Second Half	Whole Test		
100 Art Students (not advanced), Columbia University	.69	.81	14	13	24	155	2.4
100 Sixth Grade Children, P. S. 165, New York	.67	.80	13	10	20	95	1.8
25 Advanced Art Students Columbia University. First checking	.66	.79	10	11	18	176	3.30
25 Advanced Art Students, Columbia University. Second checking	.67	.80	11	10	20	178	4

NOTE: The Rechecking Coefficient was found by correlating the scores of 25 Advanced Art Students. The correlation for first and second checkings as given in the table was .87.

*For a non-selected group of 100 adults the reliability coefficient was found to be .93 for the same test of 60 items. See Table 9.

lapping of the group scores and illustrates the general superiority of the specialists. The ranking of sixth grade children is also compared. The validity of the test is shown by the fact that in groups of cultured people those who are engaged in some actual work in the field rank higher than those who do not devote their time or attention to art interest or who have not been specially trained for the work. There is considerable overlapping, and in the group of advanced art students a few rank exceptionally low.

Four persons in the miscellaneous group have as high ranking as the specialists. These persons have had some art training and are professors of cultural subjects in colleges and universities.

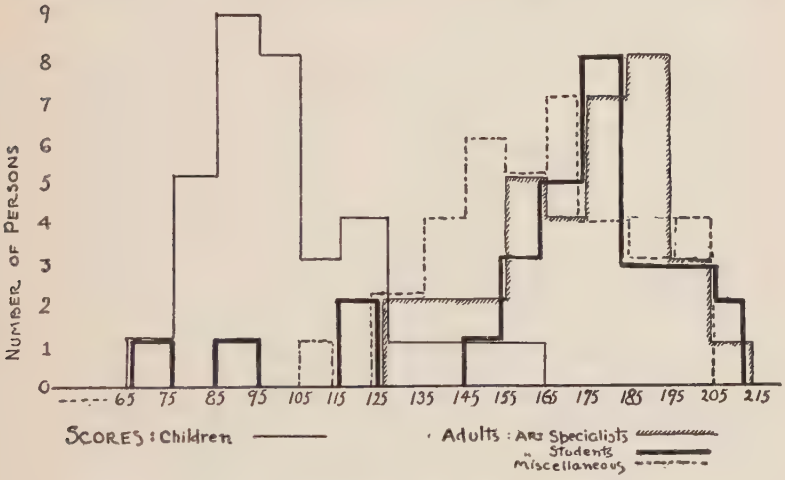

FIG. II. DISTRIBUTION OF GROUP SCORES

The range in scores of the sixth grade children is from 65 to 165. Their highest score does not quite come up to the average for the adult group. The chart shows a wide gap between the sixth grade and the groups composing the 100 judges. No comparisons have been made with data from adults of more general or "average" distribution. The early experimental tests with 2,100 children indicate that the degree of agreement with the consensus increases from grade to grade. The results are given in Table 13.

So far the reliability of the test has been studied with reference to the methods of scoring the test. The test has been shown to be valid for group comparisons inasmuch as there is an advance

from the first grade through the grades of junior high school and from the freshman year of college to graduate students and competent art critics. Table 14 gives the reliability coefficients for aptitude and character tests as a basis for comparison.

TABLE 13

COMPOSITE GRADE MEDIANS

Grade	Number of Children	Number of Tests Graded	Median Number of Correct Placements
1	306	8	4.22
2	155	8	4.26
3	332	8	5.38
4	327	8	5.56
5	265	27	15.60
6	294	27	16.17
Jr. I	218	27	17.54
Jr. II	213	27	18.22

TABLE 14

LIST OF CHARACTER AND APTITUDE TESTS

Title of Test and Source of Data*	Time Required	Number and Classification of Persons Tested	Reliability Coefficient (Rechecking)	Reliability Coefficient (Spearman-Brown Formula)	Standard Deviation
Tests in Fundamental abilities of Visual Art. Research Service Co., Los Angeles, Report Sept., 1927	55 minutes	1,100 children, Grades 3 to 9	.87	.87 ~~.85~~	Not given
Art Judgment Test, Meier-Seashore. Bureau of Educational Research and Service, University of Iowa.	45-50 minutes	2,500 children, Grades 6 to 12	~~.85~~	~~Not given~~ ~~.85~~	Not given
Mechanical Aptitudes Test, Stenquist, not manipulative skills. World Book Co., 1921	95 minutes	200 children, Grades 6, 7, 8		.6—.7	

* T. L. Kelley, *Interpretation of Educational Measurements*, World Book Co., 1927.

TABLE 14 (*Continued*)

Title of Test and Source of Data	Time Required	Number and Classification of Persons Tested	Reliability Coefficient (Rechecking)	Reliability Coefficient (Spearman-Brown Formula)	Standard Deviation
Music Test, Kwalwasser-Ruch. University of Iowa, 1924	Not given	167 children, Grades 6, 8, 10, 12		.97	51.5
Tests of Incorrigibility, Cady, V. M *Journal of Delinquency*, Monograph 2	2 hours	150 boys, 12 to 14 years of age		.746	
Character Tests, Voelker P. E. Bureau of Publications, Teachers College, Columbia University, 1921	Several hours	150 children Grades 5 to 12		.83	
I. E. R. Deception Tests, Hartshorne and May. *Studies in Deceit*, Macmillan Co.	Several hours		Parts of the Test: Copying Adding to Score Peeping Faking Lying to win Approval	 .87 .82 .72 .62 .77	
Art Test, Margaret McAdory. Given at Columbia University, Nov. 28, 1928	No limit Approximately 60 minutes	100 adults Art specialists Advanced art students People of general culture 25 art students 100 sixth grade children	 .87	.93 .80	23 .20

CHAPTER V

HOW THE TEST MAY BE USED

Educational Measurement.—The test can be used as a instrument in measuring both individual and group ability to distinguish degrees of artistic merit. This instrument of measurement is of a wide range of subjects brought together as a balanced unit by the use of single variables to test ability to distinguish merit in line and shape, dark and light, or color. Thus the test may be used to estimate abilities to judge furniture, clothing, architecture, utensils, and similar things; and at the same time an analysis can be made of the ability to distinguish merit in the separate elements of form and color on account of the division into groups according to art elements.

Careful diagnosis of this nature will suggest remedial studies. The structure of the test, including a list of subjects and art elements, has been fully described in Chapter II. The test has sufficient reliability to be successfully used in group analysis, and the ratio of the probable error of the estimated true score to the standard deviation of the scores would indicate that a test of sixty items is fairly satisfactory for individual diagnosis. Before the test was published 12 more items were added, making a total of 72 for the final form. The reliability of the test was discussed in the preceding chapter.

Norms have not been established for the test, but the further application of the test to groups of children and adults may suggest some desirable standards of attainment. At present it can only be said that the raw scores of 100 judges range from 72 to 210. This range is exhibited in Figure III. Arbitrarily, we may make the tentative statement that raw scores of 181 or more may be considered "superior"; scores of 161 through 180, "high average"; scores of 141 through 160, "low average"; and scores of 140 or below, "poor."

Median scores were found in an earlier test for large numbers of primary, elementary, and junior high school children. Table 13

shows the grade medians for grades 1 to 8. These are not to be confused with the present test but seem to indicate that tests of this type discriminate between art capacities at different grade levels.

When adequate testing has been done and standards of achieve-

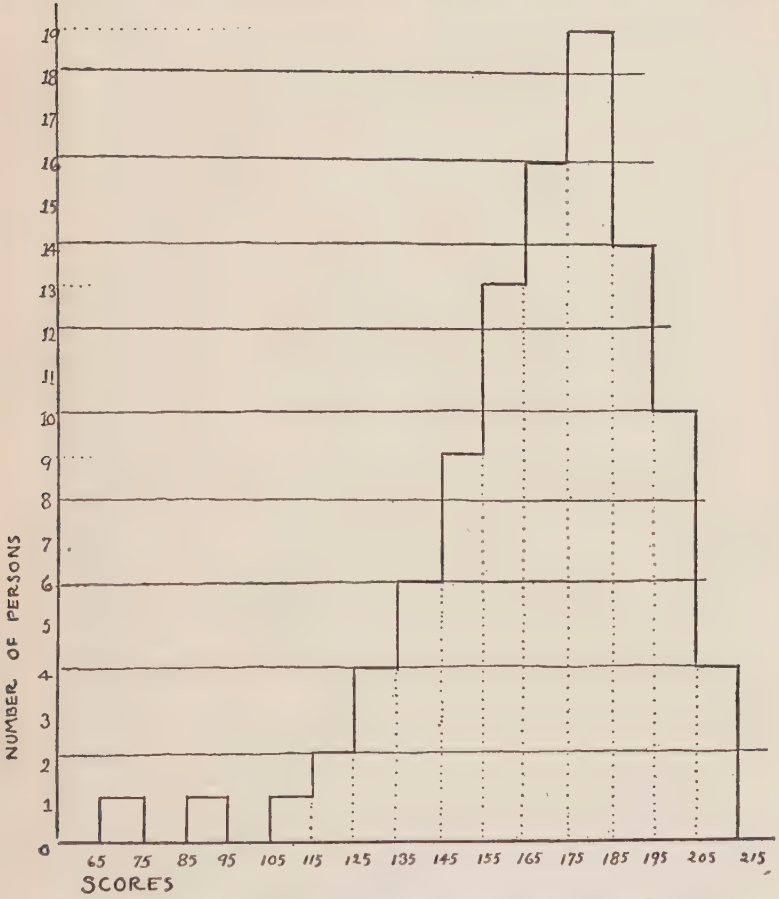

FIG. III. RANGE OF SCORES OF 100 JUDGES EMPLOYING SCORING METHOD NO. 21 OF TABLE 9

ment at different age levels and grade levels have been established, a field of study will be opened up along psychological lines for the evaluation and selection of types of subject matter suitable for use in teaching children. At present the data are insufficient to justify definite proposals on the fitting of subject matter to

grade levels, but they do suggest many lines of inquiry, such as: Is the growth of appreciation of a given art element, say form, uniform through the grades? And again, Is the growth of appreciation of one art element, say form, parallel with that of another art element, say color? Similar questions may be raised relative to classes of art subjects, such as paintings, textiles, architecture. Very carefully planned experimental study will be necessary before these facts can be known but such study would prove of great value in the field of art education. The test is so organized that facts of growth in ability to distinguish degrees of merit of the various art elements can be recognized, and it furnishes an instrument with which these psychological studies may be started.

The test items have value as teaching material. The problems requiring solution have been differentiated. Groups of tests may be taken from the whole and studied as units; for example, all the items relating to textiles and clothing or those relating to furniture or to architecture. The subject of each test item calls for æsthetic consideration. Items so constructed may be used in developing appreciation.

Uses in the Field of Production.—Many articles to be manufactured for sale might well be submitted in design to groups of competently trained people for comparison so that the greatest satisfaction may be assured for those who will purchase the article. The test suggests a method and technique of procedure. It may be of use also in the selection of persons for specialized work requiring an understanding of art value; for example, designers, art teachers, stylists, buyers, and salesmen.

Its Use to the Consumer.—In a specific way the test will bring to large groups of people the issues involved in the selection of articles for use. It will make them more conscious of differences in form and color and will enable an individual to make a fair estimate of his own ability to choose objects for his own use which he wishes to be pleasing to others also. Through analysis one will be able to tell in what respects his judgment may be deficient, and in what direction study would be helpful. The consumer may not immediately affect the production of many articles. At present, he is forced to choose from things concerning which he knows nothing of the process of production, e.g., textiles, furniture, clothing, houses, and almost everything

used. The individual ought to have within him an independence of judgment which will enable him to make his environment satisfying, whatever the whims of changing style or fashion may attempt to dictate. Eventually revolt against the production of certain types of design and decoration will turn the tide toward more satisfying forms.

CHAPTER VI

CONCLUSIONS

The research and statistical treatment of the data relating to the multiple choice art test concerned in this study gave the following findings.

First, various degrees of reliability may be found for the same multiple choice test requiring ranked order of placement, the reliability coefficients varying according to the method of scoring the test items.

Second, the higher coefficients of reliability result from giving consideration to every placement. In other words, in the art test here involved a better estimate of the ability of persons to judge artistic merit is obtained when the scores are based on the order of all the placements instead of only the first choices.

Third, of the various methods of scoring a multiple choice test such as the art studied here, the method which is most economical of time spent in scoring and which yields a high reliability is that which gives one point for each placement correctly marked. The method is simple: If the first choice is correctly placed it is given one point; if the second is correctly placed it is given one point; if the third is correctly placed it is given one point; and if the fourth is correctly placed it is given one point. Thus it is possible for four points to be credited to each item composed of four presentations which differ by approximately even steps. This method of scoring proved to be the most satisfactory from the standpoint of both economy of time and reliability of scores.

BIBLIOGRAPHY

Brooks, Fowler D. "Relative Accuracy of Ratings with and without Drawing Scales." *School and Society,* April 28, 1928.

Christensen, E. O., and Karwoski, T. "A Test in Art Appreciation." *Art Psychology Bulletin,* University of North Dakota, 1925.

Christensen, E. O. "Discrimination of Varying Degrees of Artistic Expression as a Measure of Sensitiveness in Æsthetic Reactions." Harvard University Press, 1927.

Garrett, H. E. *Statistics in Psychology and Education.* Longmans, Green & Co., 1926.

Gordon, K. "A Study of Æsthetic Judgment." *Journal of Experimental Psychology,* Feb. 1923(p. 36.

Hollingsworth, L. S. *Special Talents and Defects.* Macmillan, 1923.

Hull, Clark L. *Aptitude Testing.* World Book Co.

Jones, E. E. "The Correlation of Visual Memory and Perception of Perspective with Drawing Ability." *School and Society,* 1922, 16.

Kline, L. W., and Carey, G. L. *The Kline-Carey Measuring Scale for Free-hand Drawing.* Johns Hopkins Press, 1922.

Kelley, Truman Lee. *Interpretation of Educational Measurements.* World Book Co., 1927.

Lewerenz, Alfred S. *Tests in Fundamental Abilities of Visual Art.* Research Service Co., Los Angeles, 1927.

Meier, Norman C. "Special Artistic Talent." *Psychological Bulletin,* Vol. 25, No. 5, 1928.

Meier, Norman C. *Æsthetic Judgment as a Measure of Art Talent.* University of Iowa, 1926.

Meier-Seashore. *Art Judgment Test.* Bureau of Educational Research and Service, University of Iowa.

Pearson, Karl. *Tables for Statisticians and Biometricians.* Cambridge University Press, 1914.

Pintner, R. "Esthetics." *Psychology Bulletin,* 1920, vol. 17.

Ruch and Stoddard. *Tests and Measurements in High School Education.* World Book Co., 1927.

Rugg, Harold. *Statistical Methods Applied to Education.* Houghton Mifflin Co., 1917.

Thorndike, E. L. "Measurement of Achievement in Drawing." Teachers College, Columbia University, 1923.

Thorndike, E. L. "Tests of Æsthetic Appreciation." *Journal of Educational Psychology,* vol. 7: 509-22.

Thorndike, E. L. "Selection of Tests of Equal Difficulty by a Consensus." *Journal of Educational Research,* vol. 9: 133-39.

DATE DUE
MAY 11
MAR 16
MAY 27
AUG 18
AUG 19
SEP 3-
DARTMOUTH COLLEGE
3 3311 01072 9590

700.7
S565c

SICELOFF,M.M.
Construction & validation of an art test.

S0-GHY-209